REVISTA DE MÉTODOS CUANTITATIVOS PARA LA ECONOMÍA Y LA EMPRESA (30). Páginas 227-257.
Diciembre de 2020. ISSN: 1886-516X. D.L: SE-2927-06.
www.upo.es/revistas/index.php/RevMetCuant/article/view/3791

Predicción del nivel de cosecha de camarón blanco: el caso de una pequeña camaronera en la parroquia Tenguel del cantón Guayaquil, Ecuador

CEVALLOS-VALDIVIEZO, HOLGER
Escuela Superior Politécnica del Litoral (ESPOL) *
Correo electrónico: holgceva@espol.edu.ec

RODRÍGUEZ-CRISTIANSEN, ARIANA
Escuela Superior Politécnica del Litoral (ESPOL)
Correo electrónico: aristrod@espol.edu.ec

VALDIVIEZO-VALENZUELA, PATRICIA*
Correo electrónico: pvaldi@espol.edu.ec

ARÉVALO-AVECILLAS, DANNY
Universidad Católica de Santiago de Guayaquil (Ecuador)**
Correo electrónico: econ.darevalo@gmail.com

PADILLA-LOZANO, CARMEN**
Correo electrónico: carmen.padilla@ucsg.cu.edu.ec

RESUMEN

Actualmente el sector camaronero del Ecuador es uno de los sectores no petroleros con mayor proyección de crecimiento hacia el mercado internacional. A pesar del auge de este sector, la mayoría de los pequeños productores de camarón toman sus decisiones operativas en función del conocimiento empírico del negocio, sin considerar datos históricos ni ninguna herramienta científica como fundamento de sus decisiones. En este trabajo implementamos y comparamos técnicas de aprendizaje estadístico de vanguardia para la predicción del nivel de cosecha de camarón blanco Litopenaeus vannamei de una pequeña camaronera ubicada en la parroquia Tenguel del cantón Guayaquil, Ecuador. Datos de 35 pescas que corresponden a 7 ciclos se usaron como datos. Luego se hicieron predicciones reales de cosecha para los dos siguientes ciclos. Las técnicas comparadas son: Regresión Lineal Múltiple (RLM) por mínimos cuadrados, Árbol de Regresión CART, Bosques Aleatorios, Regresión adaptativa multivariante por tramos (MARS) y Máquinas de Soporte Vectorial (SVM). MARS sin interacciones, el modelo de RLM aditivo con selección de predictores por Best Subset Selection y SVM con Núcleo lineal produjeron un menor error de predicción por Validación Cruzada. El buen rendimiento predictivo de estos métodos fue confirmado con buenos resultados de predicción real en los dos siguientes ciclos. El uso de técnicas estadísticas de vanguardia puede ser de gran ayuda para obtener predicciones confiables, y, por tanto, para mejorar los procesos operativos de las pequeñas camaroneras.

Palabras clave: predicción, cosecha, camarón blanco Litopenaeus vannamei, aprendizaje estadístico, validación cruzada, MARS.
Clasificación JEL: C53; M11; Q22.
MSC2010: 62P20; 62G08; 62J07; 62J20.
Artículo recibido el 12 de diciembre de 2018 y aceptado el 2 de octubre de 2019.

Prediction of white shrimp harvest: the case of a small shrimp farm in Tenguel, Guayaquil-Ecuador

ABSTRACT

Shrimp sector in Ecuador is nowadays one of the fastest-growing non-oil sectors towards the international market. In despite of this growth, to our knowledge most of the little producers of shrimps in Ecuador take important operational decisions based upon empirical knowledge, without considering historical data nor any scientific tool. In this work we implement and compare state-of-the-art statistical learning techniques for the prediction of shrimp harvest (in pounds) for a little shrimp farm located in Tenguel, Guayaquil-Ecuador. For this study we used historical information collected by the farm biologist. The data was organized and put into a digital format by the authors. Data from n=35 past harvests, corresponding to 7 cycles of production, were used to train the models. We then made predictions of shrimp harvest for the next two production cycles. We compare Multiple Linear Regression by means of ordinary least squares, CART Regression Tree, Random Forests, Multivariate Adaptive Regression Splines (MARS) and Support Vector Machines (SVM). In our analysis, MARS with no interaction terms allowed, Linear Regression with best subset variable selection and SVM with linear Kernel gave the lowest prediction error estimate by Cross Validation. Their good predictive performance was confirmed with good predictions on the next two production cycles. The use of statistical techniques can be of great help to improve predictions and therefore operational processes of small shrimp farms.

Keywords: prediction, harvest, white shrimp Litopenaeus vannamei, statistical learning, cross-validation, MARS.
JEL classification: C53; M11; Q22.
MSC2010: 62P20; 62G08; 62J07; 62J20.

1. Introducción.

El sector acuícola en el Ecuador registra su comienzo a finales de los años sesenta con la explotación de salitrales y pampas. En la década de 1980 el camarón ecuatoriano vio su pico más alto desde su comienzo con un crecimiento del 600% de sus hectáreas cultivadas en comparación con la década anterior, pasando a ser considerado en el ranking mundial como el exportador número uno en el año de 1987 (Alvarado-Espinoza, 2016). Para la siguiente década, el sector había obtenido tal desarrollo que ya registraba una producción total de 250 millones de libras (Santillán-Lara, 2018). Sin embargo, a finales de la década de los 90 se registró un declive de dicho sector debido a ciertos factores exógenos, tales como: enfermedades como el síndrome de Taura y la Mancha Blanca, el fenómeno del Niño y la crisis económica del año 1999. Según los datos de la Cámara Nacional de Acuacultura, desde el año 2010 los precios del producto se han doblado, incentivando así su producción y su exportación. El camarón ecuatoriano actualmente representa el 50% de la producción latinoamericana. El Ecuador es el segundo exportador mundial de camarón, superado solamente por India (FAO, 2018). Es uno de los sectores no petroleros con mayor proyección de crecimiento hacia el mercado internacional y la industria no petrolera más tecnificada en el país.

A pesar del auge de este sector, en nuestro conocimiento la mayoría de los pequeños productores de camarón toman sus decisiones operativas teniendo en cuenta el conocimiento empírico del negocio, sin considerar datos históricos ni ninguna herramienta científica como fundamento de sus decisiones. Por ejemplo, los productores de camarón desearían predecir con cierta precisión el nivel de cosecha de camarón antes de realizar la pesca. Esto les permitiría proyectar sus ingresos y planificar de manera efectiva sus operaciones y futuras inversiones. En este contexto, los métodos estadísticos reducen la incertidumbre y ayudan a obtener una mejor predicción del nivel de cosecha en comparación con predicciones empíricas. El desarrollo de plataformas computacionales permite hoy en día la fácil implementación de técnicas complejas de aprendizaje estadístico que reducen considerablemente el error de predicción en comparación con métodos estadísticos tradicionales. Estas técnicas de aprendizaje estadístico son además apropiadas para problemas con datos con muchas variables y/u observaciones, en donde han mostrado muy buenos resultados predictivos. Por ejemplo, Garcia et al. (2007) realizó la predicción del nivel de cosecha de camarón blanco usando redes neuronales artificiales de tipo alimentación hacia adelante (feed-forward) en datos de series de tiempo en el período 1986-2004 recogidos en Charleston Harbor (Carolina del Sur, EEUU). En el modelo, los autores usaron captura por unidad de esfuerzo (CPUE), número de embarcaderos de pesca comercial estatal, salinidad y temperatura como variables explicativas del nivel de cosecha del próximo mes $(t + 1)$ y de tres meses después $(t + 3)$. Garcia et al. (2007) obtuvieron una precisión de predicción de hasta el 92% para el caso $(t + 1)$ y hasta del 79% para el caso $(t + 3)$. Sujjaviriyasup & Pitiruek (2013) compararon técnicas de aprendizaje estadístico para la predicción del nivel de cosecha de camarón blanco en datos de Tailandia en el período entre enero de 2007 y diciembre de 2012. Las siguientes técnicas fueron comparadas: ARIMA (Box & Jenkins, 2015), modelo Holt-Winters (Kalekar, 2004) y Máquinas de Soporte Vectorial (SVM por sus siglas en inglés) (Boser et al., 1992; Cortes & Vapnik, 1995; Drucker et al., 1997). En la comparación se encontró que SVM obtuvo las predicciones más precisas para este problema. Drews-Jr. et al. (2014) presentaron una nueva metodología basada en técnicas para el aprendizaje estadístico para predecir el nivel de cosecha del camarón rosado en datos del estuario de la Laguna de los Patos del estado de Río Grande del Sur (Brasil). Los datos fueron obtenidos a través de agencias gubernamentales. Los autores discretizaron la variable del nivel de cosecha y estudiaron un problema de clasificación. La nueva metodología consistía en formar una técnica de meta aprendizaje (Breiman, 1996; Vilalta & Drissi, 2002; Kuncheva, 2014) que combinaba SVM, árboles de decisión (Holte, 1993; Göndör & Bresfelean, 2012) y técnicas de aprendizaje de reglas (Cohen, 1960; Kohavi, 1995). Los autores obtuvieron una precisión de predicción de hasta el 91% con su metodología. Por otro lado, Grant et al. (1988) propuso un modelo basado en cadenas de Markov para predecir la cosecha anual de camarones en el Golfo de México. Los datos se

construyeron usando simulaciones y estaban compuestos por atributos relacionados al volumen de captura, área de pesca, profundidades y mortalidad natural y por pesca.

En este trabajo implementamos y comparamos técnicas de aprendizaje estadístico de vanguardia para la predicción del nivel de cosecha de camarón blanco *Litopenaeus vannamei* (en libras) de una pequeña camaronera ubicada en la parroquia Tenguel del cantón Guayaquil, Ecuador. Se usaron datos recopilados desde la creación de la camaronera en el mes de noviembre de 2015. En este estudio se plantea un problema de regresión. Las técnicas comparadas son: Regresión Lineal Múltiple (RLM) por medio de mínimos cuadrados (Seal, 1967; Stigler, 1981), Árbol de Regresión CART (Breiman et al., 1984), Bosques Aleatorios o Random Forests (Breiman, 2001), Regresión adaptativa multivariante por tramos o MARS (Friedman, 1991) y Máquinas de Soporte Vectorial (SVM). En nuestro conocimiento, no existen trabajos académicos hechos en el Ecuador en donde se compare rendimientos de predicción de metodologías de aprendizaje estadístico con datos locales de cosecha de camarón. Una segunda contribución de este trabajo es la identificación/selección de variables determinantes de la producción de camarón blanco. El modelo de Regresión Lineal Múltiple permite identificar variables predictoras significativas con un modelo aditivo, mientras que CART identifica variables importantes durante su construcción. La técnica de la Selección del Mejor Subconjunto o Best Subset Selection (Beale et al., 1967; Hocking & Leslie, 1967; Furnival & Wilson, 1974) permite seleccionar variables predictoras importantes para el modelo de Regresión Lineal Múltiple. Entre las técnicas consideradas, MARS es la única técnica que hace selección de variables de manera automática durante su construcción.

El resto del manuscrito está organizado de la siguiente manera. En la Sección 2 se describe la infraestructura de la camaronera, así como los procedimientos y actividades para la producción de camarón blanco en este negocio. También se explica el origen de los datos utilizados en este estudio y se da una breve explicación de las variables consideradas. En la Sección 3 se define el error de predicción teórico y se explica el procedimiento de Validación Cruzada (Lachenbruch & Mickey, 1968; Geisser, 1993) para estimar este error en el contexto de un problema de regresión. Los métodos de predicción usados en nuestro problema de regresión son explicados ampliamente en la Sección 4. La Sección 5 muestra los resultados de predicción de cada uno de los métodos implementados en base a la estimación del error por Validación Cruzada y en base a predicciones en datos futuros. Se incluye una discusión sobre los resultados de los métodos usando la descomposición del error teórico en sesgo al cuadrado, varianza y ruido. Además, se indican las variables identificadas como importantes por RLM y CART, así como las variables seleccionadas por Best Subset Selection y MARS como las más importantes para predecir el nivel de cosecha de camarón blanco. Finalmente, la Sección 6 presenta las conclusiones de este trabajo.

2. Datos.

Para la realización de este estudio se consideró la producción de 5 estanques de cultivo semi intensivo de camarón blanco *Litopenaeus vannamei* de la camaronera en estudio. Se procedió a estudiar 7 ciclos de producción de cada una de las piscinas. Estos ciclos abarcan el período comprendido entre el mes de noviembre de 2015 hasta el mes de abril de 2018. Se realizaron luego predicciones del nivel de cosecha de camarón en los dos ciclos siguientes, que corresponden al período entre mayo-agosto de 2018 y septiembre-diciembre de 2018, respectivamente.

La infraestructura de las piscinas corresponde a estanques rústicos con suelo de arcilla-arena que poseen sobre la superficie muros de contención perimetrales y divisorios con una altura aproximada de 2.50 metros. El tamaño de los estanques varía, siendo el más pequeño de 5.8 hectáreas y el más grande de 15 hectáreas. El agua usada proviene de la estación de bombeo, que trabaja con motores estacionarios y bombas de flujo axial que llevan el agua del canal hacia el reservorio y las piscinas de manera independiente por compuertas de paso de agua que por lo

general son abiertas en las noches. La salinidad del agua varía de acuerdo a la estación del año, en verano puede llegar a 20 partes por mil, mientras que en invierno a 10 partes por mil. En esta camaronera anualmente se establecen 3 ciclos de cultivo que duran en promedio 106 días cada una. En este período de tiempo se obtienen tallas de camarones promedio de 18 gramos. No existe aireación mecánica, sin embargo, se tratan de mantener los niveles óptimos de oxígeno a través del recambio de agua y del control del alimento artificial. Se utilizan fertilizantes inorgánicos y alimento artificial (balanceado de 35% de proteína) de acuerdo al tamaño del camarón. El alimento es suministrado en tres raciones diarias mediante el método del boleo, que consiste en que dos personas distribuyan ampliamente el alimento sobre el estanque empleando una canoa para una mejor movilización. La cantidad de alimento se ajusta dependiendo de los niveles de oxígeno en el agua y la revisión de la alimentación en el estanque. En los siguientes casos se suspende la alimentación en las piscinas o se disminuye la dosis suministrada de balanceado: cuando el resultado de tomar el oxígeno disuelto en la piscina a las 4:00 am es inferior a 3 mg/l, o, cuando se observan restos de balanceado en la piscina correspondientes a la dosis anterior. Si no se observa alimento restante de la dosis anterior en la piscina durante 3 días consecutivos, se aumenta el 10% de la dosis de balanceado.

Para este estudio se utilizó la información histórica recopilada por el biólogo de la camaronera. Esta información se recopiló desde el inicio de las operaciones de la camaronera en noviembre de 2015. La información fue organizada y luego digitalizada para su análisis por los autores del presente estudio. Se estudiaron datos de $n = 35$ pescas en los 5 estanques (piscinas), que corresponden a 7 ciclos o corridas. Luego se hicieron predicciones de cosecha en los dos siguientes ciclos, para cada uno de los 5 estanques. La variable de respuesta del nivel de cosecha la medimos a través del número total de libras de camarón producidas. A partir de entrevistas a expertos e investigación exploratoria se escogieron las variables que podrían tener un efecto en nuestra variable de respuesta. Todas las variables predictoras consideradas son cuantitativas. A continuación, enlistamos y describimos las variables predictoras del nivel de cosecha consideradas en este estudio:

- **Hectáreas (HAS)**: número de hectáreas que comprende el área de cada una de las piscinas.

- **Cantidad sembrada:** número de larvas que fueron depositadas en cada una de las piscinas.

- **Peso (en gramos):** peso promedio en gramos de una muestra tomada en una piscina un día antes de la pesca.

- **Días de cultivo:** número total de días que comprenden el ciclo de cultivo. Se lo obtiene con la siguiente ecuación:

  Días de cultivo = fecha de pesca - fecha de siembra

- **Supervivencia estimada (en porcentaje):** es la estimación de la supervivencia final de los camarones según la tabla propuesta por la compañía proveedora de balanceado de la camaronera bajo estudio (Tabla 1). Esta tabla utiliza como entrada una estimación empírica del nivel de mortalidad del camarón durante el período del ciclo de cultivo.

Tabla 1. Tabla de supervivencia estimada.

Nº. Días	Nº. Semanas	SUPERVIVENCIA
7	1	96%
14	2	91%
21	3	87%
28	4	82%
35	5	79%
42	6	75%
49	7	72%
56	8	69%
63	9	66%
70	10	62%
77	11	59%
84	12	56%
91	13	53%
98	14	50%
105	15	46%
112	16	43%
119	17	40%

Fuente: Compañía proveedora de balanceado de la camaronera bajo estudio.

- **Total de alimento consumido (libras)**: cantidad de alimento artificial (balanceado) suministrado a los camarones durante el ciclo de producción expresado en libras.

- **Oxígeno disuelto promedio del estanque (mg/l)**: oxígeno disuelto promedio del estanque (en mg/l) durante el ciclo de cosecha en cada una de las piscinas. Este parámetro es tomado diariamente por uno de los empleados a las 4:00 am. Su concentración debe ser la adecuada para asegurar un entorno saludable en el ciclo de la cosecha. Según Nicovita (1997), los niveles críticos de oxígeno disuelto en el agua del estanque que están relacionados directamente con el bienestar o salud del camarón son: desde 0 - 1.0 mg/l, letal; 1 - 1.5 mg/l., letal con exposición prolongada; 1.7 - 3.0 mg/l., pobre conversión alimenticia, crecimiento lento, disminución de la resistencia a las enfermedades si continúan expuestos.

- **Temperatura promedio del estanque (grados Celsius):** temperatura promedio del estanque en grados Celsius durante el ciclo de cosecha en cada una de las piscinas. Este parámetro es tomado diariamente por uno de los empleados a las 4:00 am. La temperatura del estanque influye en el crecimiento y desarrollo del camarón. Litopenaeus vannamei tiene un rango óptimo de temperatura que va de 25ºC a 30ºC, el cual es considerado adecuado para su cultivo. Sin embargo, si la temperatura cae por debajo de 25ºC o sube por encima de 30ºC la temperatura es estresante para el camarón, afectando su consumo de alimento en un 30%-50%, ya sea disminuyéndolo o aumentándolo. En estas circunstancias tampoco es aprovechado el alimento eficientemente en el crecimiento en peso (i.e. para convertirlo en músculo) (Nicovita, 1997).

La Tabla 6 de la Sección de Anexos nos muestra la tabla completa de datos recogidos y organizados para la camaronera bajo estudio.

3. El error de predicción y su estimación.

Antes de explicar la metodología que se usará para realizar la predicción del nivel de cosecha de camarón blanco, definiremos al error de predicción y explicaremos dos estimadores del error de predicción. Los estimadores del error se usan en aplicaciones prácticas para comparar la capacidad predictiva de técnicas de aprendizaje estadístico.

Los datos de la muestra en donde se ajustan los métodos de predicción se les denomina datos de entrenamiento. La calidad predictiva de un método de predicción debe de ser evaluada en datos de prueba independientes que provienen de la misma población que la muestra de entrenamiento. Esto corresponde con lo que sucede en aplicaciones prácticas en donde se hacen predicciones en datos futuros independientes a los usados para construir el método de predicción. Sea Y la variable de respuesta, $X = (X_1, X_2, \dots, X_k)$ el vector de variables predictoras y $\hat{f}(X)$ el modelo de predicción que ha sido estimado en los datos de entrenamiento. La función de pérdida que mide los errores entre Y y $\hat{f}(X)$ la denotamos como $L(Y, \hat{f}(X))$. Una opción típica para la función de pérdida en el contexto de regresión es la función de pérdida de errores cuadráticos (Hastie et al., 2001):

$$L(Y, \hat{f}(X)) = (Y - \hat{f}(X))^2$$

Con la función de pérdida de errores cuadráticos definimos al error de predicción como:

$$\text{Err} = \text{E}\left[L\left(Y, \hat{f}(X)\right)\right] = \text{E}\left[(Y - \hat{f}(X))^2\right].$$

En la práctica no es posible obtener el error de predicción ya que sólo se cuenta con una muestra de la población. Un procedimiento estándar para estimar Err es el de dividir de forma aleatoria los datos muestrales en datos de entrenamiento y datos de prueba (McLachlan, 1992). Una opción muy común es usar la regla de 80% para datos de entrenamiento y 20% para datos de prueba. Sin embargo, para nuestro estudio de predicción del nivel de cosecha de camarón en donde tenemos $n = 35$ pescas, con 28 observaciones para entrenar el método y 7 para evaluar su capacidad predictiva tendríamos una estimación muy imprecisa y con mucho sesgo del error de predicción Err. En muestras pequeñas como la de nuestro estudio, la técnica de Validación Cruzada puede obtener una estimación más estable del error de predicción (Molinaro et al., 2005; Green & Ohlsson, 2007).

Validación Cruzada (CV)

La Validación Cruzada es una técnica para estimar el error de predicción (Lachenbruch & Mickey, 1968; Geisser, 1993). El arte de la Validación Cruzada radica en que el analista sólo utiliza los datos de la muestra para obtener una estimación del error de predicción. Es decir, CV no requiere que el analista disponga de datos de prueba adicionales.

Específicamente, CV divide de manera aleatoria los datos de la muestra en P partes aproximadamente iguales. Luego, un subconjunto de $P - 1$ partes es usado para entrenar el modelo y la parte restante es usada para evaluar el modelo construido, simulando así una evaluación del modelo usando datos de prueba. Este proceso continúa hasta que todos los subconjuntos de $P - 1$ partes hayan sido usados para entrenar y la parte restante para evaluar. Al final, tendremos P estimaciones del error, que se combinan para obtener una estimación final.

Formalmente, sea $\kappa: (1, \cdots, n) \mapsto \{1, \cdots, P\}$ una función de indexación que indica la partición a la cual la i-ésima observación es designada por la aleatorización. Denotamos por $\hat{f}^{-p}(x)$ la función ajustada, obtenida sin la p-ésima parte de los datos $(p = 1, \cdots, P)$. Entonces la estimación del error por Validación Cruzada se define como:

$$CV_{error}(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}^{-p(i)}(x_i)\right)^2 \quad (1)$$

Los valores más comunes para P son $P = 5$, $P = 10$ y $P = n$. A esta última opción se la conoce también como Validación Cruzada dejando uno fuera o *leave-one-out cross-validation* (LOOCV).

Descomposición del error de predicción

Es posible descomponer el error de predicción teórico en términos del sesgo, varianza y ruido irreducible. Asumiendo que $Y = f(X) + \varepsilon$, en donde $\mathrm{E}[\varepsilon] = 0$ y $Var(\varepsilon) = \sigma_\varepsilon^2$, Geman et al. (1992) muestra que el error de predicción de una técnica de regresión ajustada $\hat{f}(X)$ en un punto $X = x_0$, usando la función de pérdida de errores cuadráticos, se descompone de la siguiente manera:

$$\mathrm{Err}(x_0) = \mathrm{E}\left[\left(Y - \hat{f}(x_0)\right)^2 | X = x_0\right]$$

$$= \sigma_\varepsilon^2 + \left[\mathrm{E}\hat{f}(x_0) - f(x_0)\right]^2 + \mathrm{E}\left[\hat{f}(x_0) - \mathrm{E}\hat{f}(x_0)\right]^2$$

El primer término representa el ruido irreducible de Y alrededor de su verdadera media $f(x_0)$. El segundo término representa el sesgo al cuadrado de $\hat{f}(x_0)$, mientras que el tercer término representa la varianza de predicción de $\hat{f}(x_0)$. Típicamente, mientras más complejo o flexible sea el modelo $\hat{f}$ construido, más pequeño será el sesgo pero más alta será la varianza. Para minimizar el error de predicción necesitamos un método de aprendizaje que logre una compensación óptima entre sesgo y varianza.

4. Metodología.

Para efectuar la predicción del nivel de cosecha de camarón se ajustarán los siguientes métodos:

4.1. Modelo de Regresión Lineal Múltiple (con errores normales).

Utiliza un modelo lineal en los parámetros para predecir una variable de respuesta cuantitativa (Y) en base a una o varias variables predictoras (X) cuantitativas o cualitativas. El modelo de Regresión Lineal Múltiple con errores normales se define de la siguiente manera:

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_k x_{i,k} + \varepsilon_i \quad (2)$$

$$i = 1, \ldots, n$$

$$\varepsilon_i \sim^{\mathrm{iid}} N(0, \sigma^2)$$

donde Y es la variable dependiente que se quiere predecir, $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ son $k + 1$ parámetros desconocidos del modelo que representan al intercepto y a las pendientes respectivamente, $x_1, x_2, \ldots, x_k$ son constantes que representan a las k variables predictoras o covariables, mientras que ε representa al término del error que contiene los otros factores distintos de $x_1, x_2, \ldots, x_k$ que afectan a Y. El subíndice i se refiere a la i-ésima observación y n es el número de observaciones. En el modelo de Regresión Lineal Múltiple con errores normales se asume que los errores ε_i son variables aleatorias independientes e idénticamente distribuidas normales, con media cero y varianza σ^2 constante (i.e. varianza constante para cualquier nivel de las variables predictoras). Esto implica además que Y_i es una variable aleatoria normal:

$$Y_i \sim N(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_k x_{i,k}, \sigma^2)$$

Con el fin de obtener un modelo parsimonioso y estable para hacer predicciones, en este estudio consideramos el modelo de regresión lineal de primer orden (i.e. lineal en los predictores) mostrado en la ecuación (2).

Estimación por Mínimos Cuadrados

En este estudio estimamos los parámetros del modelo en (2) usando mínimos cuadrados (Seal, 1967; Stigler, 1981). Sean $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ los estimadores por mínimos cuadrados de los parámetros del modelo en (2). Usando los datos disponibles, obtenemos estimaciones de estos estimadores $\mathrm{b}_0, \mathrm{b}_1, \ldots, \mathrm{b}_k$. Los valores ajustados o predicciones del modelo de acuerdo a las estimaciones por mínimos cuadrados se obtienen de la siguiente manera:

$$\hat{y} = \mathrm{b}_0 + \mathrm{b}_1 x_1 + \mathrm{b}_2 x_2 + \cdots + \mathrm{b}_k x_k, \tag{3}$$

El método de mínimos cuadrados estima la función de regresión del modelo minimizando la suma cuadrática de los residuos. En otras palabras, las estimaciones $\mathrm{b}_0, \mathrm{b}_1, \ldots, \mathrm{b}_k$ en (3) se obtienen de manera simultánea minimizando la suma cuadrática de los residuos con respecto a $\beta_0, \beta_1, \ldots, \beta_k$:

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i,1} - \cdots - \beta_k x_{i,k})^2 \tag{4}$$

Podemos expresar la solución de mínimos cuadrados en forma matricial. Para aquello, primero definimos las siguientes matrices:

$$\mathbf{Y}_{(n\times 1)} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad \mathbf{X}_{(n\times(k+1))} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1,k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2,k} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n,k} \end{bmatrix}$$

$$\boldsymbol{\beta}_{((k+1)\times 1)} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \qquad \boldsymbol{\varepsilon}_{(n\times 1)} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

de tal forma que podemos expresar el modelo en (2) de forma matricial:

$$\mathbf{Y}_{(n\times 1)} = \mathbf{X}_{(n\times(k+1))}\boldsymbol{\beta}_{((k+1)\times 1)} + \boldsymbol{\varepsilon}_{(n\times 1)}$$

en donde:

- $\mathbf{Y}$ es un vector aleatorio de respuestas
- $\boldsymbol{\beta}$ es un vector de parámetros
- $\mathbf{X}$ es una matriz de constantes
- $\boldsymbol{\varepsilon}$ es un vector de variables aleatorias normales independientes con valor esperado $\mathbf{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$ y matriz de varianzas y covarianzas:

$$\boldsymbol{\sigma}^2[\boldsymbol{\varepsilon}] = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{I}.$$

Denotamos al vector aleatorio con los estimadores por mínimos cuadrados $\hat{\beta}_0, \hat{\beta}_1, \ldots \hat{\beta}_k$ como **B**:

$$\mathbf{B}_{((k+1)\times 1)} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}$$

Entonces, los estimadores por mínimos cuadrados se definen de la siguiente forma:

$$\mathbf{B} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}$$

Procedimientos inferenciales tales como intervalos de confianza y pruebas de hipótesis pueden llevarse a cabo para los componentes de $\boldsymbol{\beta}$ y para σ^2. Detalles técnicos sobre estos procedimientos pueden encontrarse en Hastie et al. (2001) y Kutner et al. (2004), por ejemplo. Note que el supuesto de la media cero de los errores $\mathrm{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$ garantiza la insesgadez de los estimadores por mínimos cuadrados, mientras que el supuesto de varianza constante σ^2 (homocedasticidad) garantiza que los estimadores por mínimos cuadrados tengan la mínima varianza de entre todos los estimadores lineales insesgados (David & Neyman, 1938; Plackett 1949). Sin embargo, la precisión de los estimadores por mínimos cuadrados puede disminuir ante el problema de multicolinealidad (Salmerón-Gómez & Rodríguez-Martínez, 2017) o ante la inclusión de predictores innecesarios, lo que afecta también la calidad predictiva del modelo (Mundfrom et al., 2018). En varias aplicaciones, es posible mejorar la predicción haciendo una selección de predictores importantes. Al hacer esto sacrificamos un poco de sesgo para reducir la varianza de predicción y, por tanto, podríamos también reducir el error de predicción (Tibshirani, 1996; Hastie et al., 2001).

Selección de variables predictoras por el Mejor Subconjunto (Best Subset Selection)

Best Subset Selection es una técnica muy popular para hacer selección de predictores importantes en el modelo de regresión lineal. La selección por el Mejor Subconjunto encuentra para cada $s \in \{0,1,\cdots,k\}$, el subconjunto de tamaño s con la menor suma cuadrática de sus residuos en (4). El problema de obtener un tamaño óptimo s implica compensación entre sesgo y varianza (Hastie et al., 2001). Además, frecuentemente se desea ajustar un modelo pequeño que sea fácil de interpretar. Típicamente, se escoge el modelo con el menor número de predictores que minimice un estimador del error de predicción (p. ej. Validación Cruzada, ver Sección 3).

Evaluación de los supuestos del Modelo de Regresión Lineal Múltiple

El modelo que se define en (2) implica los supuestos de linealidad ($\mathrm{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$), varianza constante σ^2 y normalidad e independencia de los errores ε_i $(i = 1,\cdots,n)$. Los procedimientos usuales para construir intervalos de confianza y pruebas de hipótesis sobre los parámetros β_j $(j = 0,1,\cdots,k)$ y σ^2se desarrollan bajo estos supuestos y, por tanto, la validez de estos procedimientos inferenciales se basan en la validez de estos supuestos. El no cumplimiento de estos supuestos puede afectar también la calidad predictiva del modelo de RLM (Kutner et al., 2004). En este trabajo evaluamos los supuestos usando procedimientos estándar. Para evaluar el supuesto de linealidad se usó el gráfico de residuos versus valores ajustados, para el de homocedasticidad se usó un gráfico de residuos cuadráticos versus valores ajustados, para el supuesto de normalidad de los errores se usó el gráfico cuantil-cuantil de los residuos y para evaluar el supuesto de errores independientes se usó el gráfico de la función de autocorrelación estimada (Cook & Weisberg, 1982; Atkinson, 1985; Cook 1998). Se realizó también la prueba global para los cuatro supuestos desarrollada por Peña & Slate (2006), que controla la probabilidad de error Tipo I (que se incrementaría al combinar pruebas para cada uno de los supuestos) y que puede ser usada para detectar la violación de al menos un supuesto. Otro problema común al ajustar un modelo de

regresión lineal múltiple es el de multicolinealidad. Para detectar la presencia de multicolinealidad en nuestro problema usamos el Factor de Inflación de Varianza (VIF por sus siglas en inglés) para cada variable predictora. El VIF refleja el grado al cual la varianza muestral de los $\hat{\beta}_j$ ($j = 1, \cdots, k$) se incrementa como consecuencia de correlaciones entre las variables predictoras (Marquardt, 1970; Theil, 1971; Fox, 1984). Cuando las variables predictoras no están linealmente correlacionadas, entonces en cada variable el VIF alcanza su valor mínimo de 1. Cuando existen correlaciones el VIF en cada variable es mayor que 1, alcanzando una cantidad sin límite para una variable cuando esta posee una asociación lineal perfecta con las otras variables predictoras en el modelo.

4.2. Métodos Basados en Árboles.

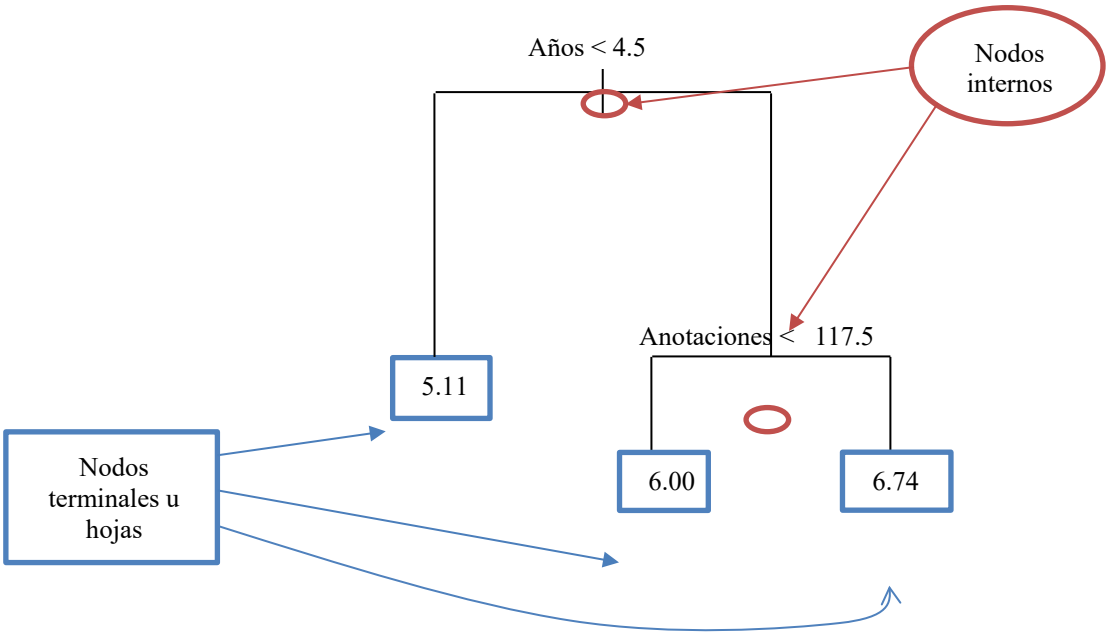
4.2.1. Árbol de Regresión (CART)

Construye un modelo no lineal (Breiman et al., 1984). CART realiza particiones binarias recursivas dentro del espacio predictor, de tal manera que las particiones resultantes sean lo más homogéneas posibles en el sentido que minimizan la suma cuadrática de los residuos en cada paso. En cada región CART estima el valor esperado condicional de Y tomando la media de los valores de Y correspondientes a las observaciones dentro de la región. El proceso de optimización para encontrar una partición es sobre todas las variables predictoras $X_1, \cdots, X_k$ y sobre todos los puntos posibles t_k de cada variable. Las particiones binarias son representadas en forma de un árbol. La Ilustración 1 nos muestra tal representación para los datos "Hitters". Este conjunto de datos contiene marcas y salarios de jugadores de béisbol de la Liga Mayor de EEUU y fueron usados en James et al. (2014). El objetivo de estudio es predecir el salario de un jugador de béisbol teniendo en cuenta sus marcas de juego alcanzadas.

Como observamos en la Figura 1, la partición de cada nodo interno es determinada por una variable predictora X_j y un punto en esta variable t_j. La partición se la efectúa de la siguiente manera: las observaciones que cumplen con la condición $X_j < t_j$ son enviadas al lado izquierdo del nodo mientras que aquellas que no la cumplen son enviadas al lado derecho. Para nuestro ejemplo con los datos "Hitters", la primera partición se la realiza con la variable predictora "Años" y su valor 4.5. Las observaciones que cumplen con esta condición son enviadas a la región del lado izquierdo y aquellas que no cumplen son enviadas a la región del lado derecho. Hemos entonces formado dos nuevas regiones o nodos. CART buscará nuevamente en estas regiones hacer una partición. En la región resultante del lado derecho se hace una partición usando la variable "Anotaciones" y su valor 117.5. Mientras que en la región del lado izquierdo no se hace más particiones. ¿Hasta cuándo CART realiza particiones? Un criterio muy estándar es el de hacer crecer el árbol hasta que haya un número mínimo de observaciones en cada nodo o hasta que quede una sola observación en el nodo.

Al hacer predicciones con un árbol muy grande corremos el riesgo de que el modelo se sobreajuste a las observaciones usadas para construirlo y que, por tanto, obtengamos una mala predicción en observaciones nuevas, i.e. observaciones a predecir en el futuro. En otras palabras, con un árbol grande se podría tener un sesgo pequeño pero una varianza de predicción mucho más grande, lo que afectaría el error de predicción. Es muy común entonces podar el árbol y encontrar un tamaño ideal para el árbol que compense entre sesgo y varianza, que son los componentes del error de predicción (ver Sección 3). Un procedimiento estándar es el de comparar varios árboles de diferentes tamaños y escoger un tamaño óptimo para el árbol que minimice un estimador del error de predicción (p. ej. Validación Cruzada, ver Sección 3). Hoy en día CART es una técnica muy popular por su flexibilidad, su fácil interpretación al identificar predictores importantes en su construcción y porque es capaz de manejar valores perdidos sin necesidad de hacer imputación previa (Feelders, 1999; Cevallos-Valdiviezo & Van Aelst, 2015).

Figura 1. Ejemplo de árbol de regresión para los datos "Hitters".



En este ejemplo se busca predecir el salario de un jugador de béisbol sobre la base de sus marcas de juego alcanzadas.

Fuente: James et al. (2014), pág. 304.

4.2.2. Bosques Aleatorios (Random Forests)

Se basa en la combinación de L árboles de decisión de tal manera que formen un bosque (Breiman, 2001). En particular, se toman varias muestras bootstrap sobre las que se ajusta un árbol en cada una. Los árboles del bosque no son árboles CART. Los árboles de los bosques durante su crecimiento seleccionan de manera aleatoria un subconjunto de variables de tamaño m del conjunto completo de predictores de tamaño k ($m \leq k$) como variables candidatas para realizar la partición en cada nodo. De este modo Bosques Aleatorios controla la correlación entre los árboles. En el contexto de regresión el método de Bosques Aleatorios obtiene la predicción final en cada observación al promediar su predicción en los L árboles individuales del bosque. De esta manera, Bosques Aleatorios obtiene predicciones más estables en comparación con el árbol simple CART (Louppe, 2014; Cevallos-Valdiviezo & Van Aelst, 2015). Para nuestro problema de predicción del nivel de cosecha de camarón usamos $m = k/3$, es decir, el número de predictores considerados en cada partición es igual al número total de predictores dividido para 3. Éste es un criterio muy estándar para m en varias aplicaciones.

4.3. Regresión adaptativa multivariante por tramos (Multivariate Adaptive Regression Splines, MARS).

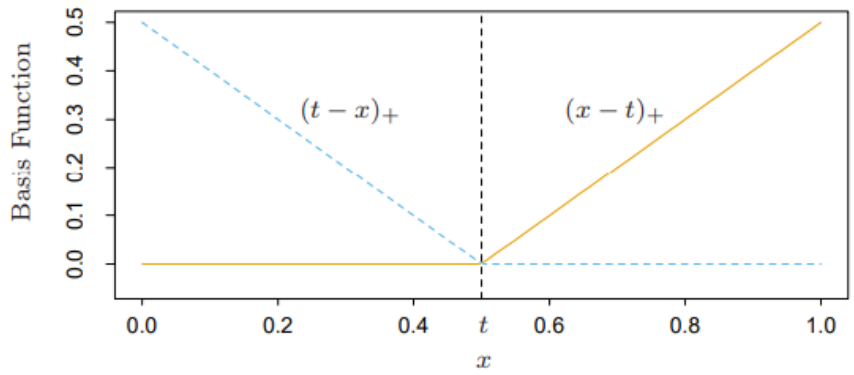
MARS es un método flexible y con estructura a la vez (Friedman, 1991). La flexibilidad de MARS es gracias al uso de funciones base. Sin embargo, MARS controla su flexibilidad e impone una estructura al ajustar funciones base lineales definidas por tramos. Estas funciones base lineales son trazadores lineales (splines) que expanden el espacio predictor. Las funciones base lineales usadas por MARS tienen la forma $(x - t)_+$ y $(t - x)_+$, donde el "+" significa la parte positiva:

$$(x - t)_+ = \begin{cases} x - t, & si\ x > t, \\ 0, & de\ otra\ manera \end{cases}$$

$$(t-x)_+ = \begin{cases} t-x, & si\ x < t, \\ 0, & de\ otra\ manera \end{cases}$$

La Figura 2 nos muestra un ejemplo de estas funciones base lineales cuando $t = 0.5$. La premisa de MARS es formar "pares reflejados" para cada variable X_j con nudos en cada valor observado x_{ij} de dicha variable.

Figura 1. Ejemplo de funciones base $(\mathbf{x}-\mathbf{t})_+$ (color naranja) y $(\mathbf{t}-\mathbf{x})_+$ (color celeste).



Fuente: Hastie et al. (2001), pág. 322.

Algoritmo

En el paso 1 del algoritmo de MARS el conjunto C contiene todos los pares reflejados, para todas las variables y para todos los valores observados:

$$C = \left\{(X_j - t)_+,\ (t - X_j)_+\right\}$$

$$t \in \{x_{1j}, \cdots, x_{nj}\}$$

$$j = 1, \ldots, k.$$

Si todos los valores observados de X_j son distintos $(j = 1, \ldots, k)$, en el paso 1 el conjunto C contiene $2nk$ funciones base en total. En cada paso del algoritmo de MARS el conjunto C es actualizado eliminando funciones base y añadiendo productos de dos o más de estas funciones. El algoritmo de MARS realiza una regresión lineal paso a paso hacia adelante con las funciones base como variables predictoras:

$$f(X) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m\,(X),$$

en donde las $h_m(X)$ son funciones base del conjunto C. Los coeficientes β_m son estimados usando mínimos cuadrados. A continuación, describimos brevemente la primera parte del algoritmo usado por MARS:

1. Empezamos con la función constante $h_0(X) = 1$ sola en el modelo y todas las demás funciones en el conjunto C son funciones candidatas a entrar en el modelo.

2. Ingresa al modelo el par reflejado que produzca la reducción máxima del error de predicción estimado. Es decir, ingresa aquel par reflejado que produzca la mejor predicción.

3. Una vez que ingresa el primer par reflejado se actualiza el conjunto C de las funciones candidatas. La actualización de C consiste en retener las funciones base que estaban en el paso 1 excepto aquel par reflejado que ingresó al modelo en el paso 2. Además, se incluye al producto de los pares reflejados que ingresaron con las funciones base que quedaron.

4. Ingresa al modelo aquel par reflejado o función base que maximice la reducción del error de predicción estimado.

5. Los siguientes pasos actualizan C de la misma manera e ingresan al modelo el par reflejado o función base que maximice la reducción del error de predicción estimado.

Al ingresar un gran número de pares reflejados o funciones base al modelo podemos construir un modelo grande de tamaño $\mathcal{M}$. Durante la construcción del modelo podemos permitir que ingresen funciones base con interacciones entre predictores o sin interacciones si queremos un modelo menos flexible. Un modelo muy grande de tamaño $\mathcal{M}$ sobreajustaría los datos disponibles, dificultando la generalización del modelo para predecir datos futuros. Por tanto, se aplica un procedimiento de eliminación hacia atrás para construir un modelo más estable. En cada paso del procedimiento de eliminación hacia atrás excluimos del modelo al par reflejado o función base que produzca un mínimo aumento del error estimado. Es decir, en cada paso eliminamos el par reflejado o función base menos importante para la predicción. Se aplica el procedimiento de eliminación hacia atrás hasta que nos quedemos únicamente con la función base $h_0(X) = 1$ dentro del modelo. Como consecuencia, tendremos una secuencia de modelos, desde el más grande de tamaño $\mathcal{M}$ hasta el más simple que contiene únicamente la función base $h_0(X) = 1$. Si nos quedamos con el modelo más grande tendremos un modelo muy complejo y flexible por lo que tendremos una varianza de predicción alta, aunque un sesgo pequeño. Si nos quedamos con el modelo más simple la varianza de predicción será baja, pero el sesgo será muy alto. Por tanto, tenemos que seleccionar el tamaño óptimo del modelo usando un estimador del error de predicción (p. ej. Validación Cruzada, ver Sección 3).

El modelo final puede contener un subconjunto de todas las variables predictoras. Además, puede incluir interacciones entre variables predictoras si permitimos ingresar funciones base con interacciones en el procedimiento hacia adelante. Por tanto, además de ser un método flexible con estructura, otra de las ventajas de MARS es que permite realizar selección de variables y de interacciones de manera automática durante su construcción.

4.4. Máquinas de Soporte Vectorial (SVM).

SVM es una técnica que puede ser formulada para problemas de clasificación y regresión. SVM fue inicialmente propuesta para problemas de clasificación en Boser et al. (1992) y en Cortes & Vapnik (1995). Luego, una formulación de SVM para regresión fue propuesta por Drucker et al. (1997). SVM es una técnica que no hace supuestos probabilísticos, por lo que se la considera como una alternativa robusta ante desviaciones de estos supuestos en los datos.

Para nuestro problema de predicción consideramos regresión por SVM. La adaptación de SVM para regresión hereda algunas de las propiedades del clasificador SVM, por lo que explicamos brevemente el problema de optimización que el clasificador SVM busca resolver. Con clases linealmente separables es posible encontrar un hiperplano separador óptimo que maximiza un margen entre las clases (i.e. que maximiza la distancia mínima entre cada clase y el hiperplano), bajo la restricción que los puntos estén correctamente clasificados y que se encuentren fuera del margen. Con clases solapadas, no existe una solución para el hiperplano separador. Para encontrar una solución con clases linealmente no separables se generaliza el problema del hiperplano separador permitiendo que algunos puntos caigan dentro del margen o

inclusive en el lado incorrecto del hiperplano (i.e. permitiendo clasificaciones erróneas). A esta generalización se la conoce como Clasificador de Soporte Vectorial. El problema del Clasificador de Soporte Vectorial puede ser formulado como un problema de optimización con penalización, en donde se penaliza soluciones con mayores errores de clasificación. La solución del Clasificador de Soporte Vectorial viene dada por los puntos sobre el margen, puntos que caen dentro del margen y por los puntos que caen en el lado incorrecto del hiperplano. Es decir, la solución viene dada por los puntos que son más difíciles de clasificar. A estos puntos se los conoce también como vectores soporte. El Clasificador de Soporte Vectorial construye por tanto un hiperplano en el espacio predictor original. Un procedimiento más flexible se obtiene al expandir el espacio predictor usando funciones base y luego ajustar en este espacio expandido el Clasificador de Soporte Vectorial. La Máquina de Soporte Vectorial (SVM) es una extensión de esta idea, en donde se permite que la dimensión del espacio predictor expandido sea muy grande o infinita. Al considerar transformaciones asociadas con funciones Núcleos (Kernels) de forma apropiada, el problema de SVM puede formularse como un problema general de regularización, cuya solución tiene dimensión finita y viene dada por una combinación lineal de n funciones Núcleo evaluadas en los datos de entrenamiento.

En el espacio predictor original, Regresión de Soporte Vectorial busca por su parte un hiperplano con una banda de ancho ϵ, bajo la restricción que esta banda contenga todos los puntos. Sin embargo, para valores pequeños de ϵ no existe solución para este hiperplano, por lo que se relaja la restricción para encontrar una solución permitiendo que un cierto número de observaciones caiga fuera de la banda. La variable de holgura ξ_i o ξ_i^* determina la distancia máxima a la que puede encontrarse una observación y_i, $i = 1, \cdots, n$, sobre o debajo de la banda respectivamente. Definimos al hiperplano por:

$$f(\mathbf{x}) = \beta_0 + \mathbf{x}^{\mathrm{t}}\boldsymbol{\beta}$$

en donde $\mathbf{x}^{\mathrm{t}} = [x_1, x_2, \ldots, x_k]$ y $\boldsymbol{\beta}^{\mathrm{t}} = [\beta_1, \beta_2, \ldots, \beta_k]$ es el vector de k variables predictoras y de k parámetros respectivamente. Note que en este caso $\boldsymbol{\beta}$ no incluye el parámetro del intercepto β_0. Regresión de Soporte Vectorial minimiza:

$$\min_{\beta_0, \boldsymbol{\beta}} \left[\frac{1}{2} \|\boldsymbol{\beta}\|^2 + \frac{1}{\lambda} \sum_{i=1}^{n} (\xi_i + \xi_i^*) \right] \tag{5}$$

sujeto a:

$$y_i - f(\mathbf{x}_i) \leq \epsilon + \xi_i$$

$$f(\mathbf{x}_i) - y_i \leq \epsilon + \xi_i^*$$

$$\xi_i \geq 0$$

$$\xi_i^* \geq 0, i = 1, \cdots, n.$$

donde $\lambda > 0$ es un parámetro de ajuste que controla el número de observaciones fuera de la banda. Se puede mostrar fácilmente que la formulación del problema en (5) es equivalente a la siguiente formulación (Hastie et al., 2001):

$$\min_{\beta_0, \boldsymbol{\beta}} \left[\sum_{i=1}^{n} L\big(y_i, f(\mathbf{x}_i)\big) + \frac{\lambda}{2} \|\boldsymbol{\beta}\|^2 \right] \tag{6}$$

en donde $L\big(y_i, f(\mathbf{x}_i)\big) = \max(0, |y_i - f(\mathbf{x}_i)| - \epsilon)$. Note que si $L\big(y_i, f(\mathbf{x}_i)\big) = \big(y_i - f(\mathbf{x}_i)\big)^2$, i.e. la función de pérdida de errores cuadráticos, Regresión de Soporte Vectorial es equivalente a la Regresión Contraída (Hoerl & Kennard, 1970). La solución de este problema viene dada por:

$$\widehat{\boldsymbol{\beta}} = \sum_{i=1}^{n} (\hat{\alpha}_i - \hat{\alpha}_i^*)\, \mathbf{x}_i$$

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^{n} (\hat{\alpha}_i - \hat{\alpha}_i^*)\, \mathbf{x}^{\mathrm{t}} \mathbf{x}_i + \hat{\beta}_0 \qquad (7)$$

en donde $\hat{\alpha}_i$ y $\hat{\alpha}_i^*$ son la solución de la función dual, que es un problema de programación cuadrática:

$$\max_{\substack{\alpha_i, \alpha_i^* \\ i=1,\cdots,n}} \sum_{i=1}^{n} (\alpha_i - \alpha_i^*)\, y_i - \epsilon \sum_{i=1}^{n} (\alpha_i + \alpha_i^*) - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (\alpha_i - \alpha_i^*)\left(\alpha_j - \alpha_j^*\right) \mathbf{x}_i^{\mathrm{t}} \mathbf{x}_j \qquad (8)$$

sujeto a: $\sum_{i=1}^{n} (\alpha_i - \alpha_i^*) = 0$

$$0 \leq \alpha_i, \alpha_i^* \leq {}^{1}/_{\lambda}$$

$$\alpha_i \alpha_i^* = 0.$$

Debido a la naturaleza de las restricciones, sólo un subconjunto de puntos cumple con la condición $(\hat{\alpha}_i - \hat{\alpha}_i^*) \neq 0$, el cual corresponde a puntos sobre la banda o que se encuentran fuera de la banda. Estos son los vectores soporte para Regresión de Soporte Vectorial. Note que una vez encontrados $\hat{\alpha}_i$ y $\hat{\alpha}_i^*$, $\hat{\beta}_0$ se puede obtener fácilmente con cualquier vector soporte sobre la banda usando la ecuación $y_i - \hat{f}(\mathbf{x}_i) = \epsilon$ o la ecuación $\hat{f}(\mathbf{x}_i) - y_i = \epsilon$.

De manera análoga al problema de clasificación, SVM para regresión se basa en la idea de expandir el espacio predictor usando funciones base $h(\mathbf{x}) = \left(h_1(\mathbf{x}), h_2(\mathbf{x}), \cdots, h_M(\mathbf{x})\right)$ y ajustar Regresión de Soporte Vectorial en el espacio expandido. Definimos al hiperplano en el espacio expandido como:

$$f(\mathbf{x}) = \beta_0 + h(\mathbf{x})^{\mathrm{t}} \boldsymbol{\beta}.$$

Ya que el problema de optimización de Regresión de Soporte Vectorial en (8) y la solución para el hiperplano en (7) involucra a las variables predictoras a través de productos punto, podemos escribir:

$$\max_{\substack{\alpha_i, \alpha_i^* \\ i=1,\cdots,n}} \sum_{i=1}^{n} (\alpha_i - \alpha_i^*)\, y_i + \epsilon \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (\alpha_i - \alpha_i^*)\left(\alpha_j - \alpha_j^*\right) h(\mathbf{x}_i)^{\mathrm{t}} h(\mathbf{x}_j) \qquad (9)$$

con solución:

$$\hat{f}(\mathbf{x}) = \hat{\beta}_0 + \sum_{i=1}^{n} (\hat{\alpha}_i - \hat{\alpha}_i^*)\, h(\mathbf{x})^{\mathrm{t}} h(\mathbf{x}_i)$$

Una formulación equivalente a (9), puede obtenerse al considerar un problema de estimación funcional en un espacio Hilbert con Núcleo reproductor $\mathcal{H}_K$ generado por un Núcleo positivo definido K. Consideremos funciones base de transformación h asociadas con funciones base ϕ_m de una expansión (posiblemente finita) de K:

$$K(\mathbf{x}, \mathbf{x}') = \sum_{m=1}^{\infty} \phi_m(\mathbf{x})\, \phi_m(\mathbf{x}') \delta_m$$

$$\delta_m \geq 0$$

$$\sum_{m=1}^{\infty} \delta_m^2 < \infty$$

con $h_m(\mathbf{x}) = \sqrt{\delta_m}\phi_m(\mathbf{x})$. Note que $h(\mathbf{x})^{\mathrm{t}} h(\mathbf{x}') = \sum_{m=1}^{\infty} \phi_m(\mathbf{x})\, \phi_m(\mathbf{x}')\delta_m = K(\mathbf{x}, \mathbf{x}')$.

Por tanto, no es necesario especificar la transformación $h(\mathbf{x})$, sólo se requiere conocer la función Núcleo $K(\mathbf{x}, \mathbf{x}')$. Ya que $f(\mathbf{x}) = \beta_0 + \sum_{m=1}^{\infty} \beta_m \sqrt{\delta_m}\, \phi_m(\mathbf{x})$ y tomando $\theta_m = \sqrt{\delta_m}\beta_m$, podemos escribir (6) como:

$$\min_{\substack{\beta_0, \theta_m \\ m=1,\cdots,\infty}} \left[\sum_{i=1}^{n} L(y_i, \beta_0 + \sum_{m=1}^{\infty} \theta_m\, \phi_m(\mathbf{x}_i)) + \frac{\lambda}{2} \sum_{m=1}^{\infty} \frac{\theta_m^2}{\delta_m}\right] \tag{10}$$

que es un problema de dimensión infinita. Sin embargo, se puede mostrar que la solución a este problema es de dimensión finita (Wahba, 1990) y se expresa como:

$$\hat{f}(\mathbf{x}) = \hat{\beta}_0 + \sum_{i=1}^{n} (\hat{\alpha}_i - \hat{\alpha}_i^*)\, K(\mathbf{x}, \mathbf{x}_i) \tag{11}$$

i.e. una combinación lineal de n funciones Núcleo evaluadas en cualquier punto $\mathbf{x}$ y en cada uno de los puntos de entrenamiento $\mathbf{x}_i$, $i = 1, \cdots, n$. El criterio de optimización en (10) es por tanto equivalente al criterio en (9). Usando (11) y propiedades de $\mathcal{H}_K$, el criterio en (10) puede reducirse al criterio de dimensión finita:

$$\min_{\boldsymbol{\alpha}} [\sum_{i=1}^{n} L(\boldsymbol{y}, \mathbf{K}\tilde{\boldsymbol{\alpha}}) + \lambda\, \tilde{\boldsymbol{\alpha}}^{\mathrm{t}} \mathbf{K} \tilde{\boldsymbol{\alpha}}] \tag{12}$$

en donde $\mathbf{K}$ es una matriz de dimensión $(n \times n)$ con entrada (i, j)-ésima $K(\mathbf{x}_i, \mathbf{x}_j)$ y $\tilde{\boldsymbol{\alpha}}$ es un vector columna de dimensión $(n \times 1)$ que contiene los coeficientes $(\alpha_i - \alpha_i^*)$, $i = 1, \cdots, n$. Note que por simplicidad hemos asumido en (12) que $\beta_0 = 0$. El criterio en (12) puede verse como un caso especial de un problema general de regularización:

$$\min_{f \in \mathcal{H}} \left[\sum_{i=1}^{n} L(y_i, f(\mathbf{x}_i)) + \lambda J(f)\right]$$

donde $L(y_i, f(\mathbf{x}_i))$ es una función de pérdida, $J(f)$ es el término de penalización funcional y $\mathcal{H}$ es un espacio de funciones en donde $J(f)$ se encuentra definido.

Entre las funciones Núcleo más populares tenemos: funciones de base radiales, función Núcleo lineal, polinomial, Laplaciano y tangente hiperbólica. Asimismo, entre las funciones de pérdida más populares para regresión por SVM tenemos: ϵ-insensible, función de pérdida de Huber y la clásica función de pérdida de errores cuadráticos. Además del parámetro de penalización λ, regresión por SVM debe de optimizar el parámetro ϵ que controla el ancho de la banda. Adicionalmente, la función Núcleo a menudo depende de al menos un parámetro de ajuste, p. ej. el ancho de banda en las funciones de bases radiales. Esto implica que regresión por SVM busca optimizar sus parámetros de ajuste en un espacio de al menos dos dimensiones.

5. Resultados y Discusión.

En esta sección comparamos la capacidad predictiva de cinco metodologías para el problema de la camaronera bajo estudio. A saber: Regresión Lineal Múltiple (RLM), Árbol de Regresión (CART), Bosques Aleatorios (Random Forests), Regresión adaptativa multivariante por tramos (MARS) y Regresión por Máquinas de Soporte Vectorial (SVM). Para evaluar la capacidad predictiva de cada método, consideramos la estimación del error según la técnica de Validación Cruzada con $P = 10$ y la predicción real de cosechas futuras. Para la evaluación con Validación Cruzada se consideraron los datos de todas las cosechas observadas desde noviembre de 2015 hasta abril de 2018 (35 pescas en total). Estos datos constituyen los datos de entrenamiento. Luego se hicieron predicciones de la cosecha a recogerse en los 5 estanques durante los dos siguientes ciclos, cuyos períodos comprendían los meses de mayo-agosto de 2018 y septiembre-diciembre de 2018, respectivamente. Una vez efectuada la pesca, estimamos el error de predicción al comparar el nivel de cosecha real en cada estanque con el nivel de cosecha predicho por la metodología. La implementación de estos métodos es hecha en el lenguaje estadístico R (R Core Team, 2018). La Tabla 4 reporta para cada método la raíz cuadrada del error por Validación Cruzada CV_{error} definido en (1), así como la función y el paquete correspondiente usados para su implementación en R. La Tabla 5 reporta el error de predicción estimado en los dos ciclos siguientes para cada método.

Note que cada método tiene parámetros de ajuste que pueden usarse para mejorar la predicción. Estos parámetros son optimizados en los datos de entrenamiento. En R, estos parámetros de ajuste se controlan en los argumentos de la función respectiva. La columna "Argumentos de la función en R" de la Tabla 4 muestra algunos de los parámetros de ajuste usados en cada técnica. Regresión lineal Múltiple tiene como parámetro de ajuste el número de predictores en el modelo. Para nuestro análisis consideramos el modelo aditivo con todos los 8 predictores y usando un subconjunto óptimo de predictores a través de la técnica Best Subset Selection con búsqueda exhaustiva (i.e. búsqueda del mejor subconjunto recorriendo el espacio entero de subconjuntos). En base al Criterio de Información Bayesiano (BIC, por sus siglas en inglés), Best Subset Selection seleccionó un modelo con 4 predictores como óptimo. A saber, el modelo con los predictores: "Cantidad sembrada", "Peso (en gramos)", "Supervivencia Estimada" y "Total de alimento consumido (libras)". El modelo seleccionado por Best Subset Selection tuvo mejores resultados de predicción que el modelo completo según se muestra en la Tabla 4 y en la Tabla 5. Esto implica que en el modelo completo existen variables ruidosas o redundantes, y que por ende, no aportan información nueva para la predicción. Podemos constatar en los resultados inferenciales del modelo completo (ver Tabla 2) que al 5% de significancia cuatro variables tienen efectos no significativos. Note además que los valores VIF para el modelo completo son mayores que estos valores para el modelo reducido (ver Tabla 2 y Tabla 3, respectivamente), lo que es un indicativo de mayor multicolinealidad en el modelo completo. De hecho, las variables con efectos significativos en el modelo completo concuerdan con las variables seleccionadas por Best Subset Selection (note que la variable "Total de alimento consumido (libras)" es marginalmente significativa en el modelo completo). En cuanto a los supuestos, los gráficos de diagnóstico muestran una ligera desviación del supuesto de homocedasticidad tanto en el modelo completo como en el modelo reducido (ver Figura 4 y Figura 5 en la Sección de Anexos). Sin embargo, la prueba global concluye que todos los cuatro supuestos son aceptables al 5% de significancia en cada modelo.

Tabla 2. Valores VIF, coeficientes estimados [IC 95%] y valores P para el modelo de RLM aditivo completo con 8 predictores.

Predictor	VIF	Coeficiente de regresión [IC 95%]	Valor P
Hectáreas (HAS)	3.59	273.92 [-200.59; 748.42]	0.25
Cantidad sembrada	4.65	0.01 [0.01; 0.02]	<0.001
Peso (en gramos)	1.46	1295.16 [800.58; 1789.73]	<0.001
Días de cultivo	1.95	10.13 [-72.77; 93.02]	0.80
Supervivencia Estimada (en porcentaje)	3.57	37902.08 [29149.93; 46654.23]	<0.001
Total de alimento consumido (libras)	4.43	0.07 [-0.01; 0.14]	0.05
Oxígeno disuelto promedio del estanque (mg/l)	2.67	67.57 [-2585.29; 2450.15]	0.96
Temperatura promedio del estanque (grados Celsius)	1.21	65.69 [-697.94; 829.31]	0.86
Resultados Prueba Global: H_o: Los 4 supuestos (linealidad, homocedasticidad, errores normales e independientes) se cumplen H_a: Al menos uno de los cuatro supuestos no se cumple Valor Estadístico = 8.91 Valor P = 0.06			

En la parte inferior se muestran los resultados de la Prueba Global para evaluar los cuatro supuestos.
Fuente: Elaboración propia.

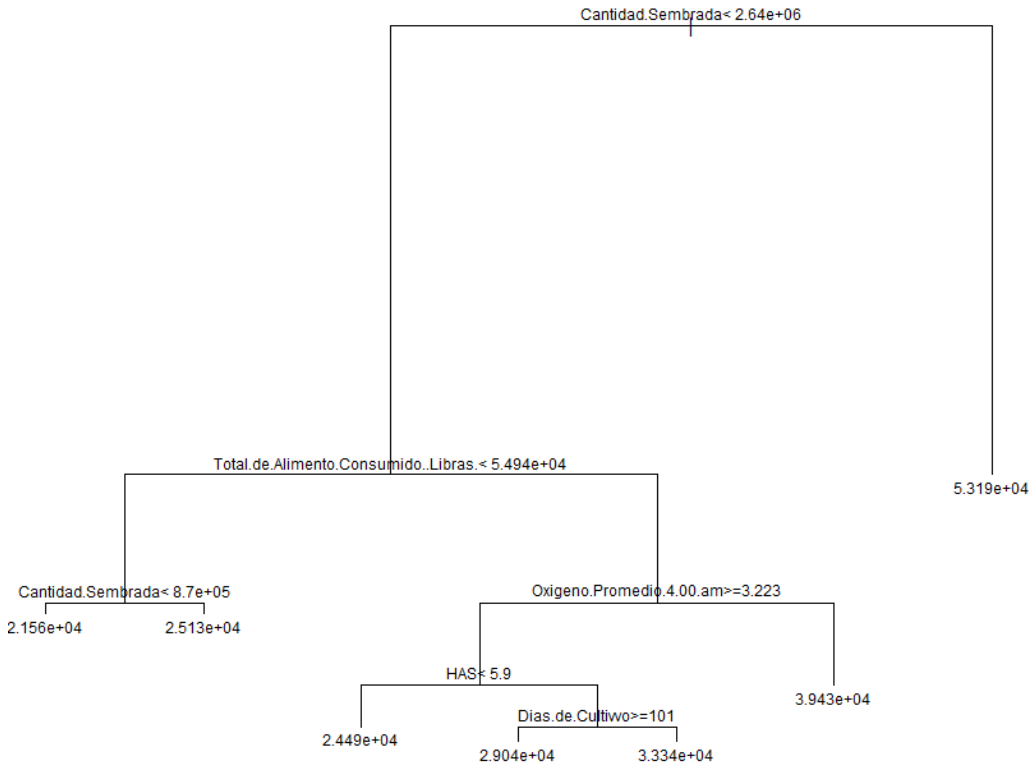
Tabla 3. Valores VIF, coeficientes estimados [IC 95%] y valores P para el modelo de RLM aditivo con 4 predictores seleccionados por Best Subset Selection.

Predictor	VIF	Coeficiente de regresión [IC 95%]	Valor P
Cantidad sembrada	3.27	0.01 [0.01; 0.02]	<0.001
Peso (en gramos)	1.39	1216.34 [750.62; 1682.05]	<0.001
Supervivencia Estimada (en porcentaje)	2.98	37821.35 [30096.66; 45546.04]	<0.001
Total de alimento consumido (libras)	1.57	0.08 [0.04; 0.12]	<0.001
Resultados Prueba Global: H_o: Los 4 supuestos (linealidad, homocedasticidad, errores normales e independientes) se cumplen H_a: Al menos uno de los cuatro supuestos no se cumple Valor Estadístico = 8.60 Valor P = 0.07			

En la parte inferior se muestran los resultados de la Prueba Global para evaluar los cuatro supuestos.
Fuente: Elaboración propia.

Árbol de regresión (CART) controla su flexibilidad a través de su tamaño. Un árbol CART más grande generalmente construye un procedimiento más flexible. Entre los parámetros de ajuste que controlan el tamaño del árbol CART tenemos: el número mínimo de observaciones en cualquier nodo terminal (argumento minbucket en rpart), el número mínimo de observaciones que deben existir en un nodo para intentar hacer una partición (argumento minsplit en rpart) y el número de nodos finales (controlado por el argumento cp en rpart). Para nuestro análisis consideramos minsplit=20 y minbucket=7. El argumento cp representa al parámetro de complejidad de CART. Mientras más grande sea el valor del argumento cp, en general más pequeño será el árbol construido. Para nuestro análisis usamos el valor default de cp=0.01 que construye un árbol con 7 nodos (ver Figura 3). Podar el árbol no mejoró el desempeño predictivo de CART, por lo que no mostramos estos resultados en esta Sección.

Figura 2. Árbol de regresión con 7 nodos terminales para predecir la cosecha de camarón blanco en nuestro estudio.



Fuente: Elaboración propia, usando los datos de entrenamiento.

La técnica de Bosques Aleatorios tiene como parámetros de ajuste al número L de árboles en el bosque y al número m de predictores seleccionados al azar que serán considerados como candidatos al hacer una partición en cada árbol individual. Usando Validación Cruzada LOOCV obtuvimos un valor óptimo de $L = 500$. Dado $L = 500$, el valor $m = 3$ dio los mejores resultados predictivos.

Para MARS consideramos modelos que no permitían interacción entre predictores, modelos hasta con interacciones de primer orden y modelos hasta con interacciones de segundo orden. En esta Sección sólo reportamos los resultados del modelo MARS sin interacciones

(degree=1 en la función earth en R), que tuvo un menor error de predicción por Validación Cruzada y mejor desempeño en la predicción de cosechas futuras.

Para regresión por SVM se usó la función de pérdida ϵ-insensible y se consideraron varias opciones para funciones Núcleo con diferentes valores para sus parámetros de ajuste, diferentes valores para el parámetro de penalización λ y para el parámetro del ancho de banda ϵ. Se consideraron las funciones Núcleo lineal, polinomial, Laplaciano, tangente hiperbólica y funciones de base radial. La opción de parámetros de ajuste con mejor desempeño predictivo consistió en un Núcleo lineal (kernel="vanilladot" en la función ksvm en R), con parámetro de penalización C=10 (donde C tiene correspondencia con el parámetro λ) y $\epsilon = 0.1$. Esta selección de parámetros para regresión por SVM es la que se muestra en las Tablas 4 y 5, y es la que se reporta en esta Sección.

Sobre la base de la estimación del error por Validación Cruzada, concluimos que el modelo MARS sin interacciones es aquél que tiene la mejor capacidad predictiva, seguido por el modelo de Regresión lineal múltiple con el subconjunto óptimo de 4 predictores. RLM es un modelo bastante estructurado en el sentido que ajusta un hiperplano para hacer la predicción. Además, RLM con selección de variables importantes a través de Best Subset Selection tiene la capacidad de eliminar variables ruidosas o redundantes, construyendo de esa forma un modelo aún más estable. MARS es más flexible ya que usa transformaciones no lineales en el espacio predictor, aunque mantiene una estructura en el modelo al usar funciones base lineal. La técnica de regresión por SVM con Núcleo lineal también tuvo un buen desempeño relativo según Validación Cruzada. Estos resultados sugieren que para obtener una buena predicción del nivel de cosecha en nuestro problema es importante construir un modelo lo suficientemente estructurado y estable. Por otro lado, la técnica de Validación Cruzada sugiere que la flexibilidad de CART y Bosques Aleatorios no ayudan a mejorar la predicción. De igual modo, SVM con funciones Núcleo más complejas que la lineal no mostraron un buen desempeño predictivo.

Tabla 4. Métodos Predictivos, funciones y su paquete correspondiente en R, artículo de referencia, argumentos de la función utilizados y error por CV.

Método Predictivo	**Función en R**	**Paquete en R**	**Referencia**	**Argumentos de la función en R**	$\sqrt{CV_{error}}$
RLM	lm()	stats	R Core Team (2018)	formula*	3508.34
Best subset selection (4 predictores)	regsubsets()	leaps	Lumley & Miller (2017)	formula.bss**, method="exhaustive"	2555.02
Árbol de Regresión (CART)	rpart()	rpart	Therneau & Atkinson (2018)	formula*, cp=0.01, control=rpart.control(minsplit=20, minbucket=7)	7989.21
Bosques aleatorios	randomForest()	randomForest	Liaw & Wiener (2002)	formula*, $m = 3$, $L = 500$	6606.83
MARS	earth()	earth	Milborrow et al. (2018)	formula*, degree=1	2.330.78

Regresión SVM	ksvm()	kernlab	Karatzoglou et al. (2004)	formula*, kernel="vanilladot", C=10, epsilon=0.1	2877.05

*formula = Hectáreas (HAS) + Cantidad sembrada + Peso (en gramos) + Días de cultivo + Supervivencia Estimada (en porcentaje) + Total de alimento consumido (libras) + Oxígeno disuelto promedio del estanque (mg/l) + Temperatura promedio del estanque (grados Celsius)

** formula.bss = Cantidad sembrada + Peso (en gramos) + Supervivencia Estimada (en porcentaje) + Total de alimento consumido (libras)

Fuente: Elaboración propia.

Además de estimar el error de predicción de cada método usando la técnica de Validación Cruzada, se evaluó su capacidad predictiva real en los dos siguientes ciclos. Para realizar esta predicción se utilizaron los valores observados en las variables predictoras durante estos dos siguientes ciclos. Estimamos el error de predicción de cada método tras obtener el monto real de la cosecha en cada uno de los 5 estanques. La Tabla 5 muestra la raíz cuadrada del error cuadrático promedio de predicción en los dos siguientes ciclos para cada uno de los métodos descritos en la Tabla 4. La Tabla 5 muestra que los desempeños predictivos reales de los métodos van en concordancia con las estimaciones del error obtenidas con Validación Cruzada. Estos resultados confirman que los modelos que son suficientemente estructurados y estables tienen un buen desempeño predictivo en nuestro problema.

Tabla 5. Raíz cuadrada del error cuadrático promedio de predicción en los dos siguientes ciclos para cada uno de los métodos considerados.

Método Predictivo	Raíz cuadrada del error cuadrático promedio
RLM	6733.84
Best subset selection (4 predictores)	6587.96
Árbol de Regresión (CART)	10575.74
Bosques aleatorios	7135.97
MARS	4016.45
Regresión SVM	6456.50

Fuente: Elaboración propia.

Otras de las contribuciones de este trabajo es la identificación/selección de variables importantes para la predicción del nivel de cosecha de camarón blanco. El modelo de RLM completo identificó a cuatro predictores significativos, los cuales concordaron con la selección óptima de Best Subset Selection. Estos predictores son:

- Cantidad sembrada (número de larvas)
- Peso (en gramos)

- Supervivencia estimada (en porcentaje)

- Total de alimento consumido (libras)

Según se visualiza en la Figura 3, los predictores identificados como los más importantes para la predicción del nivel de cosecha de camarón según CART son:

- Cantidad sembrada (número de larvas)

- Total de alimento consumido (libras)

- Oxígeno disuelto promedio del estanque (mg/l)

- Hectáreas (HAS)

- Días de cultivo

MARS realiza selección automática de predictores importantes. Para nuestro problema MARS seleccionó:

- Cantidad sembrada (número de larvas)

- Peso (en gramos)

- Sobrevivencia estimada (en porcentaje)

- Días de cultivo

- Total de alimento consumido (libras)

Note que las variables "Cantidad Sembrada" y "Total de Alimento Consumido (libras)" fueron identificadas por RLM y CART y seleccionadas por MARS y Best Subset Selection como variables importantes para la predicción del nivel de cosecha de camarón blanco.

6. Conclusiones.

Hemos realizado una comparación de cinco metodologías de aprendizaje estadístico para la predicción del nivel de cosecha de camarón blanco *Litopenaeus vannamei* (en libras) de una pequeña camaronera ubicada en la parroquia Tenguel del cantón Guayaquil, Ecuador. Para este efecto, se estudiaron datos de $n = 35$ pescas que corresponden a 7 ciclos, los cuales se usaron como datos de entrenamiento. Luego se hicieron predicciones reales de cosecha para los dos siguientes ciclos. MARS sin interacciones, el modelo de RLM aditivo con selección de predictores por Best Subset Selection y SVM con Núcleo lineal produjeron un menor error de predicción por Validación Cruzada en comparación con los métodos CART y Bosques Aleatorios. El buen rendimiento predictivo de estos métodos fue confirmado con buenos resultados de predicción real en los dos siguientes ciclos. Las variables "Cantidad Sembrada" y "Total de Alimento Consumido (libras)" fueron identificadas/seleccionadas por los métodos considerados como variables importantes para la predicción del nivel de cosecha de camarón blanco.

Se espera que este trabajo produzca un impacto positivo en la pequeña camaronera estudiada. El uso de técnicas estadísticas de vanguardia puede ser de gran ayuda para obtener predicciones confiables de la próxima cosecha. Si la camaronera en estudio puede obtener predicciones confiables, entonces podrá proyectar sus ingresos y planificar de manera efectiva

sus operaciones y futuras inversiones. A lo largo de la última década se ha comprobado que la tecnificación de procesos y la tecnología ha sido pieza clave para el crecimiento del sector camaronero en el Ecuador.

Referencias

Alvarado-Espinoza, F. (2016). *La comercialización del camarón ecuatoriano en el mercado internacional y su incidencia en la generación de divisas*. (Tesis de fin de Máster). Universidad de Guayaquil, Guayaquil.

Atkinson, A.C. (1985). *Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis.* Oxford: Clarendon Press.

Beale, E.M.L., Kendall, M.G., & Mann, D.W. (1967). The discarding of variables in multivariate analysis. *Biometrika*, *54*(3/4), 357-366.

Boser, B.E., Guyon, I.M., & Vapnik, V.N. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the Fifth Annual Workshop on Computational Learning Theory COLT '92*, 144-152.

Box, G.E.P., Jenkins, G.M., Reinsel, G.C., & Ljung, G.M. (2015). *Time series analysis: Forecasting and control* (5ta ed.). Hoboken, New Jersey: John Wiley & Sons.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*, 123-140.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5-32.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees*. Nueva York: Wadsworth & Brooks.

Cevallos-Valdiviezo, H., & Van Aelst, S. (2015). Tree-based prediction on incomplete data using imputation or surrogate decisions. *Information Sciences*, *311*, 163-181.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37-46.

Cook, R.D. (1998). *Regression Graphics: Ideas for Studying Regressions through Graphics*. Nueva York: John Wiley & Sons.

Cook, R.D., & Weisberg, S. (1982). *Residuals and Influence in Regression*. Nueva York: Chapman & Hall.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273-297.

David, F.N., & Neyman, J. (1938). Extension of the Markoff's theorem on least squares. *Statistical Research Memoirs*, *2*, 105-116.

Drews-Jr, P., Bauer, M., Machado, K., Puciarelli, P., & Felipe Dumont, L. (2014, octubre). A machine learning approach to predict the pink shrimp harvest in the patos lagoon estuary. *KDMILE*. Sao Carlos, Brasil.

Drucker, H., Burges, C.J.C., Kaufman, L., Smola, A.J., & Vapnik, V. (1997). Support vector regression machines. *Advances in Neural Information Processing Systems*, *28*(7), 155-161.

FAO (2018). *GLOBEFISH Highlights: A Quarterly Update on World Seafood Markets* (1st issue). Descargado de http://www.fao.org/3/I8626EN/i8626en.pdf

Feelders, A. (1999). Handling missing data in trees: Surrogate splits or statistical imputation? *Principles of Data Mining and Knowledge Discovery* (pp. 329-334). Berlin Heidelberg: Springer.

Fox, J. (1984). *Linear Statistical Models and Related Methods, with Applications to Social Research*. Nueva York: John Wiley.

Friedman, J.H. (1991). Multivariate adaptive regression splines. *Annals of Statistics*, *19*(1), 1-67.

Furnival, G.M., & Wilson, R.W. (1974). Regressions by leaps and bounds. *Technometrics*, *16*(4), 499-511.

Garcia, S.P., DeLancey, L.B., Almeida, J., & Chapman, R. (2007). Ecoforecasting in real time for commercial fisheries: The Atlantic white shrimp as a case study. *Marine Biology*, *152*, 15-24.

Geisser, S. (1993). *Predictive Inference*. Nueva York: Chapman & Hall.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, *4*, 1-58.

Grant, W., Matis, J., & Miller, W. (1988). Forecasting commercial harvest of marine shrimp using a Markov chain model. *Ecological Modelling*, *43*(3), 183-193.

Green, M., & Ohlsson, M. (2007, julio). Comparison of standard resampling methods for performance estimation of artificial neural network ensembles. *Third International Conference on Computational Intelligence in Medicine and Healthcare*. Plymouth, Reino Unido.

Göndör, M., & Bresfelean V. (2012). RepTree and M5P for measuring fiscal policy influences on the Romanian capital market during 2003-2010. *International Journal of Mathematics and Computers in Stimulation*, *4*, 378-386.

Hastie, T., Tibshirani, R., & Friedman, J.H. (2001). *The Elements of Statistical Learning.* Nueva York: Springer.

Hocking, R. R., & Leslie, R. N. (1967). Selection of the best subset in regression analysis. *Technometrics*, *9*(4), 531-540.

Hoerl, A.E., & Kennard, R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55-67.

Holte, R.C. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, *11*(1), 63-90.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2014). *An Introduction to Statistical Learning: With Applications in R*. Nueva York: Springer.

Kalekar, P. S. (2004). *Time series Forecasting using Holt-Winters Exponential Smoothing*. Kanwal Rekhi School of Information Technology. Descargado de https://caohock24.files.wordpress.com/2012/11/04329008_exponentialsmoothing.pdf

Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). kernlab - an S4 package for kernel methods in R. *Journal of Statistical Software*, *11*(9), 1-20.

Kohavi, R. (1995). The power of decision tables. *European Conference on Machine Learning (ECML)*, 174-189.

Kuncheva, L. I. (2014). *Combining Pattern Classifiers: Methods and Algorithms* (2da ed.). Nueva York: John Wiley.

Kutner, M., Nachtsheim, C., Neter, J., & Li, W. (2004). *Applied Linear Statistical Models*. Chicago: McGraw-Hill.

Lachenbruch, P.A., & Mickey, M.R. (1968). Estimation of error rates in discriminant analysis. *Technometrics*, *10*(1), 1-11.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomforest. *R News*, *2*(3), 18-22.

Louppe, G. (2014). *Understanding Random Forests: From Theory to Practice*. (Tesis doctoral no publicada). Universidad de Lieja, Lieja.

Lumley, T., & Miller, A. (2017). *leaps: Regression Subset Selection*. R package version 3.0.

Marquardt, D.W. (1970). Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics*, *12*(3), 591-612.

McLachlan, G. (1992). *Discriminant analysis and statistical pattern recognition.* Nueva York: John Wiley.

Milborrow, S., Hastie, T., Tibshirani, R., Miller, A., & Lumley, T. (2018). *earth: Multivariate Adaptive Regression Splines*. R package version 4.6.3.

Molinaro, A.M., Simon, R., & Pfeiffer, R.M. (2005). Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, *21*(15), 3301-3307.

Mundfrom, D., Smith, M., & Kay, L. (2018). The effect of multicollinearity on prediction in regression models. *General Linear Model Journal*, *44*, 24-28.

Nicovita (1997). *Interrelaciones de la temperatura, oxígeno y amoniaco tóxico en el cultivo de camarón en Tumbes*. Descargado de https://www.nicovita.com.pe/extranet/Boletines/ago_97_02.pdf

Peña, E.A., & Slate, E. H. (2006). Global validation of linear model assumptions. *Journal of the American Statistical Association*, *101*(473), 341-354.

Plackett, R. L. (1949). A historical note on the method of least squares. *Biometrika*, *36*(3/4), 458-460.

R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Salmerón-Gómez, R., & Rodríguez-Martínez, E. (2017). Métodos cuantitativos para un modelo de regresión lineal con multicolinealidad. Aplicación a rendimientos de letras del tesoro. *Revista de Métodos Cuantitativos para la Economía y la Empresa*, *24*, 169-189.

Santillán-Lara, X. (2018). *La acuacultura del camarón y su impacto sobre el ecosistema de manglar*. SPINCAM 3. Descargado de http://www.spincam3.net/data/actividades/2018/marzo/INFORME_TALLER_ECOSISTEMAS_USO_PRESIONES_28MAR2018.pdf

Seal, H.L. (1967). Studies in the history of probability and statistics. xv: The historical development of the gauss linear model. *Biometrika*, *54*(1/2), 1-24.

Stigler, S.M. (1981). Gauss and the invention of least squares. *The Annals of Statistics*, *9*(3), 465-474.

Sujjaviriyasup, T., & Pitiruek, K. (2013). Agricultural product forecasting using machine learning approach. *International Journal of Mathematical Analysis*, *7*(38), 1869-1875.

Theil, H. (1971). *Principles of Econometrics*. Nueva York: John Wiley.

Therneau, T., & Atkinson, B. (2018). *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-13.

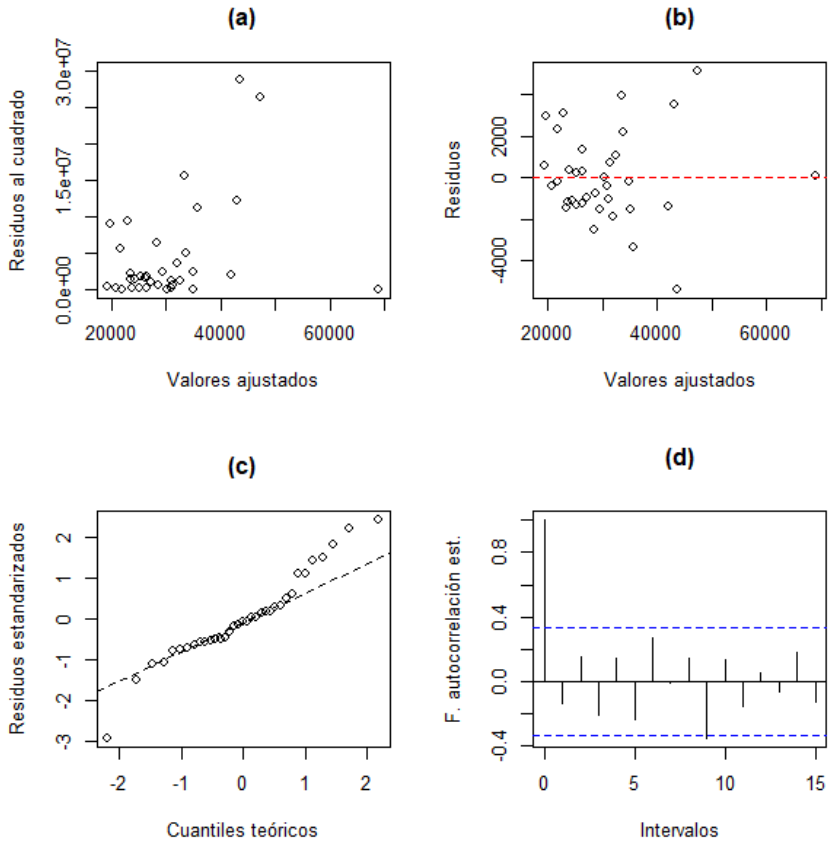
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267-288.

Vilalta, R., & Drissi, Y. (2002). A perspective view and survey of meta-learning. *Artificial Intelligence Review*, *18*(2), 77-95.

Wahba, G. (1990). *Spline Models for Observational Data*. Montpelier: Capital City Press.

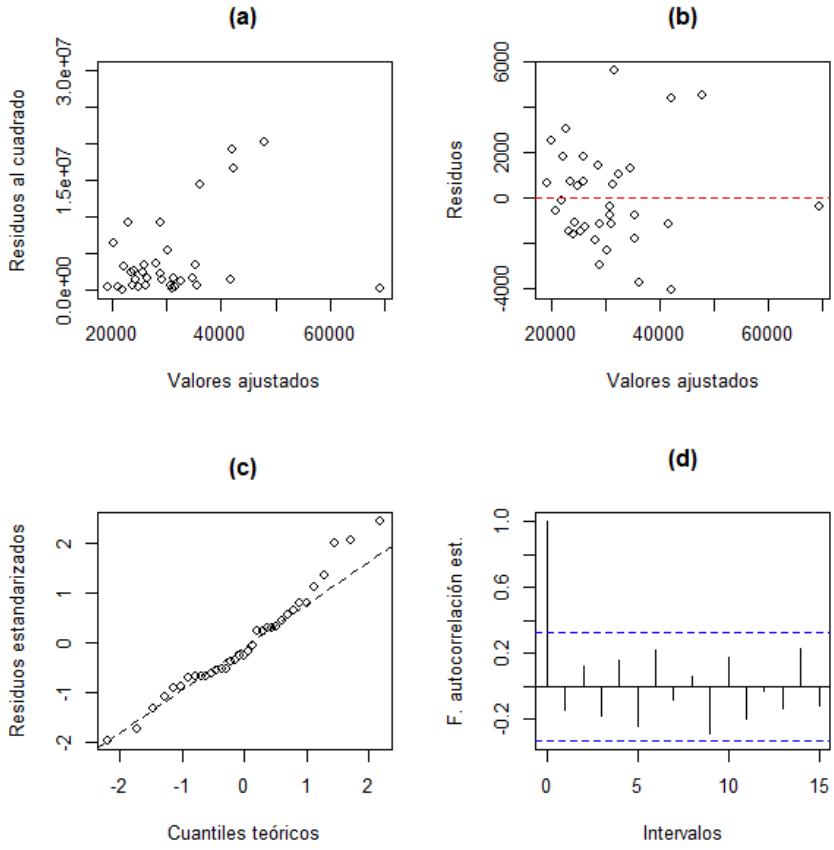
Anexos

Figura 4. Gráficos de diagnóstico para evaluar los supuestos del modelo de RLM completo con 8 predictores: (a) Residuos al cuadrado vs Valores ajustados, (b) Residuos vs Valores ajustados, (c) Residuos estandarizados vs Cuantiles teóricos de la distribución normal estándar, (d) Función de autocorrelación estimada vs Intervalos.



Fuente: Elaboración propia, usando los datos de entrenamiento.

Figura 5. Gráficos de diagnóstico para evaluar los supuestos del modelo de RLM con 4 predictores seleccionados por Best Subset Selection: (a) Residuos al cuadrado vs Valores ajustados, (b) Residuos vs Valores ajustados, (c) Residuos estandarizados vs Cuantiles teóricos de la distribución normal estándar, (d) Función de autocorrelación estimada vs Intervalos.



Fuente: Elaboración propia, usando los datos de entrenamiento.

Tabla 6. Producción de la camaronera bajo estudio desde noviembre/2015 a diciembre/2018.

Año	Corrida	Piscina	Hectáreas (Has)	Fecha de inicio	Cantidad Sembrada	Peso Promedio (Gramos)	Fecha Pesca	Días de Cultivo	Libras Cosechadas	Cliente	Supervivencia Estimada (Porcentaje)	Total de Alimento Consumido (Libras)	Oxígeno Disuelto Promedio (Miligramo por litro)	Temperatura Promedio (Grados Celsius)
2016	Primera Corrida	1	7,4	16/11/2015	900.000	17,2	23/2/2016	99	24.230	Expalsa	49,1%	39.017,00	4,21	28,41
2016	Primera Corrida	2	6,0	16/11/2015	840.000	17,5	7/3/2016	112	21.700	Expalsa	43,2%	42.636,00	4,28	28,46
2016	Primera Corrida	3	5,8	16/11/2015	840.000	18,0	5/3/2016	110	19.835	Promarisco	44,1%	37.054,60	3,57	28,46
2016	Primera Corrida	4	8,1	16/11/2015	1.020.000	18,0	6/3/2016	111	27.750	Expalsa	43,6%	46.420,00	4,50	28,64
2016	Primera Corrida	5	15,0	28/12/2015	2.004.000	13,5	26/4/2016	120	30.000	Expalsa	39,5%	86.229,00	3,63	29,29
2016	Segunda Corrida	1	7,4	4/3/2016	900.000	18,0	9/6/2016	97	25.388	Expalsa	50,1%	42.627,20	4,51	28,81
2016	Segunda Corrida	2	6,0	2/4/2016	840.000	17,5	20/7/2016	109	24.000	Expalsa	44,6%	54.940,60	4,22	28,20
2016	Segunda Corrida	3	5,8	2/4/2016	840.000	16,0	22/7/2016	111	20.350	Expalsa	43,63%	46.063,60	4,35	28,53
2016	Segunda Corrida	4	8,1	2/4/2016	1.020.000	13,0	6/7/2016	95	25.000	Expalsa	51,0%	46.755,50	4,23	28,67
2016	Segunda Corrida	5	15,0	2/4/2016	2.004.000	17,0	24/6/2016	83	30.200	Expalsa	56,5%	56.205,60	4,36	28,88
2016	Tercera Corrida	1	7,4	22/6/2016	1.750.000	18,0	3/11/2016	134	27.900	Expalsa	39,0%	103.554,00	3,82	26,48
2016	Tercera Corrida	2	6,0	3/8/2016	840.000	17,1	17/11/2016	106	25.800	Cristiansen	45,9%	66.968,00	3,60	26,21
2016	Tercera Corrida	3	5,8	1/8/2016	840.000	18,0	15/11/2016	106	22.400	Proceos	45,9%	60.651,80	3,52	26,21
2016	Tercera Corrida	4	8,1	27/7/2016	1.020.000	19,0	12/11/2016	108	29.945	Expalsa	45,0%	88.682,00	3,26	26,29
2016	Tercera Corrida	5	15,0	10/7/2016	1.980.000	17,0	2/11/2016	115	40.500	Expalsa	41,8%	151.998,00	3,18	26,36
2017	Primera Corrida	1	7,4	13/12/2016	1.015.000	18,0	11/3/2017	88	32.000	Expalsa	54,2%	64.460,00	3,40	28,36
2017	Primera Corrida	2	6,0	13/12/2016	884.000	17,9	15/3/2017	92	33.500	Cristiansen	52,4%	58.443,00	3,35	28,45
2017	Primera Corrida	3	5,8	13/12/2016	885.000	18,0	12/3/2017	89	27.875	Expalsa	53,7%	58.509,00	3,34	28,38
2017	Primera Corrida	4	8,1	13/12/2016	1.075.000	17,5	13/3/2017	90	33.600	Expalsa	53,3%	68.233,00	3,63	28,57
2017	Primera Corrida	5	15,0	13/12/2016	2.035.000	20,0	13/3/2017	90	46.500	Expalsa	53,3%	116.281,00	3,12	28,56
2017	Segunda Corrida	1	7,4	1/4/2017	1.150.000	22,1	10/8/2017	131	26.705	Cristiansen	39,0%	60.368,00	4,52	26,54
2017	Segunda Corrida	2	6,0	13/5/2017	1.200.000	18,0	12/8/2017	91	22.620	Omarsa	52,8%	41.514,00	4,36	26,19
2017	Segunda Corrida	3	5,8	1/4/2017	900.000	22,2	26/7/2017	116	23.205	Omarsa	41,3%	58.982,00	3,88	26,44
2017	Segunda Corrida	4	8,1	13/5/2017	1.500.000	17,5	24/8/2017	103	25.897	Omarsa	47,3%	43.120,00	5,00	26,29
2017	Segunda Corrida	5	15,0	5/5/2017	2.800.000	22,1	11/8/2017	98	38.160	Omarsa	49,6%	63.195,00	4,51	26,43
2017	Tercera Corrida	1	7,4	7/9/2017	800.000	16,0	8/12/2017	92	21.892	Omarsa	52,4%	42.020,00	4,29	25,67
2017	Tercera Corrida	2	6,0	7/9/2017	1.050.000	15,8	10/12/2017	94	23.992	Omarsa	51,4%	42.130,00	3,99	25,65
2018	Tercera Corrida	3	5,8	11/9/2017	2.000.000	18,4	2/2/2018	144	35.961	Cristiansen	39,0%	89.144,00	3,02	26,61
2017	Tercera Corrida	4	8,1	7/9/2017	1.025.000	15,5	27/12/2017	111	30.508	Omarsa	43,6%	57.530,00	3,90	25,68
2017	Tercera Corrida	5	15,0	7/9/2017	1.400.000	17,2	12/12/2017	96	37.400	Omarsa	50,5%	54.945,00	4,12	25,75

Año	Corrida	Piscina	Hectáreas (Has)	Fecha de inicio	Cantidad Sembrada	Peso Promedio (Gramos)	Fecha Pesca	Días de Cultivo	Libras Cosechadas	Cliente	Supervivenci Estimada (Porcentaje)	Total de Alimento Consumido (Libras)	Oxígeno Disuelto Promedio (Miligramo por litro)	Temperatura Promedio (Grados Celsius)
2018	Primera Corrida	1	7,4	18/12/2017	2.480.000	20,7	4/4/2018	107	32.410	Omarsa	45,5%	62.970,60	3,36	28,62
2018	Primera Corrida	2	6,0	18/12/2017	2.350.000	20,2	21/4/2018	124	34.740	Cristiansen	39,0%	63.823,10	2,94	28,46
2018	Primera Corrida	3	5,8	18/12/2017	2.350.000	16,8	19/3/2018	91	26.195	Omarsa	52,8%	37.753,10	3,10	28,30
2018	Primera Corrida	4	8,1	18/12/2017	2.800.000	20,7	22/4/2018	125	52.465	Proexpo	39,0%	94.342,60	3,07	28,77
2018	Primera Corrida	5	15,0	19/12/2017	4.500.000	16,1	8/4/2018	110	68.940	Cristiansen	44,1%	118.525,00	3,07	28,84
2018	Segunda Corrida	1	7,4	18/4/2018	2.200.000	18,1	12/8/2018	116	21.970	Proexpo	25,0%	43.584,20	3,36	28,62
2018	Segunda Corrida	2	6,0	5/5/2018	1.550.000	15,1	7/9/2018	125	23.695	Proexpo	46,1%	43.320,20	2,94	28,46
2018	Segunda Corrida	3	5,8	18/4/2018	1.800.000	17,0	17/8/2018	121	28.600	Omarsa	42,4%	47.196,60	3,10	28,30
2018	Segunda Corrida	4	8,1	5/5/2018	1.950.000	14,6	6/9/2018	124	22.250	Proexpo	35,4%	52.470,00	3,07	28,77
2018	Segunda Corrida	5	15,0	18/4/2018	4.000.000	17,5	10/9/2018	145	50.096	Proexpo	32,5%	98.879,00	3,07	28,84
2018	Tercera Corrida	1	7,4	31/8/2018	1.800.000	15,6	17/12/2018	108	21.280	Cristiansen	34,3%	32.551,21	4,14	26,35
2018	Tercera Corrida	2	6,0	26/9/2018	1.920.000	16,5	10/1/2019	106	20.765	Proexpo	29,8%	30.000,47	3,95	25,81
2018	Tercera Corrida	3	5,8	31/8/2018	1.600.000	18,0	8/1/2019	130	32.972	Proexpo	51,9%	49.969,92	3,81	26,18
2018	Tercera Corrida	4	8,1	26/9/2018	2.400.000	16,8	22/1/2019	118	33.152	Proexpo	37,4%	50.532,10	4,02	25,99
2018	Tercera Corrida	5	15,0	26/9/2018	4.800.000	17,4	7/2/2019	134	47.685	Proexpo	26,0%	93.453,84	3,89	26,18

Fuente: Elaboración propia.